MECHANIZED SEARCHING

Searching for recorded facts needed to solve problems in chemical technology is often a tedious and time-consuming task. Laborious effort also is often involved in the correlation of facts scattered through a number of technical papers, patents, laboratory notebooks, reports, or other private files. During recent years, various mechanical devices—especially punched cards—have proved their value as tools for both searching and correlating chemical facts and theories (5). Accomplishments in this direction have served as the basis for a continuing development program aimed at providing powerful new means for utilizing recorded facts in the chemical field (17).

Although concerned principally with the searching and correlation of chemical literature, this article also considers—though more briefly—how the accessory operations of recording, storing, reproducing, and circulating information have been facilitated by such techniques as microfilm, microcards, and "Xerography."

The surprising effectiveness of punched cards and other devices as aids to searching and correlating finds its explanation in the nature of chemical information itself. Consider, for example, any chemical product. We are, of course, interested in its chemical composition and structure. But beyond this, we will also probably be interested in methods and raw materials used for making the product, its physical and chemical properties, and its uses. Information on a chemical reaction is of interest with regard not only to reactants and reaction products, but also to conditions influencing the course and rate of the reaction, accompanying phenomena, apparatus used, etc. In general, then, chemical information requires analysis in terms of numerous independent factors.

In the usual alphabetical or numerical file the cards must be kept in order according to some system. It is possible to choose a system which would place the cards in a meaningful order. An example would be arranging compounds according to melting point and arranging references according to author or date of publication. However, such a file could be kept in only one of these arrangements without making an additional set of cards for each additional arrangement. On the other hand, a punched-card file, properly coded, can be searched mechanically according to any of the categories for which it is coded, or for any combination of categories. Also, at will, the whole file can be arranged in order according to any category.

Searches directed to combinations of subjects may serve additional purposes, one of which is studying relationships between factors. Particularly important in this connection is the use of mechanical searching as an aid in establishing correlations, for example, cause and effect relationships (12). A well-organized punched-card file can serve as an effective tool for evaluating the significance of recorded data in planning

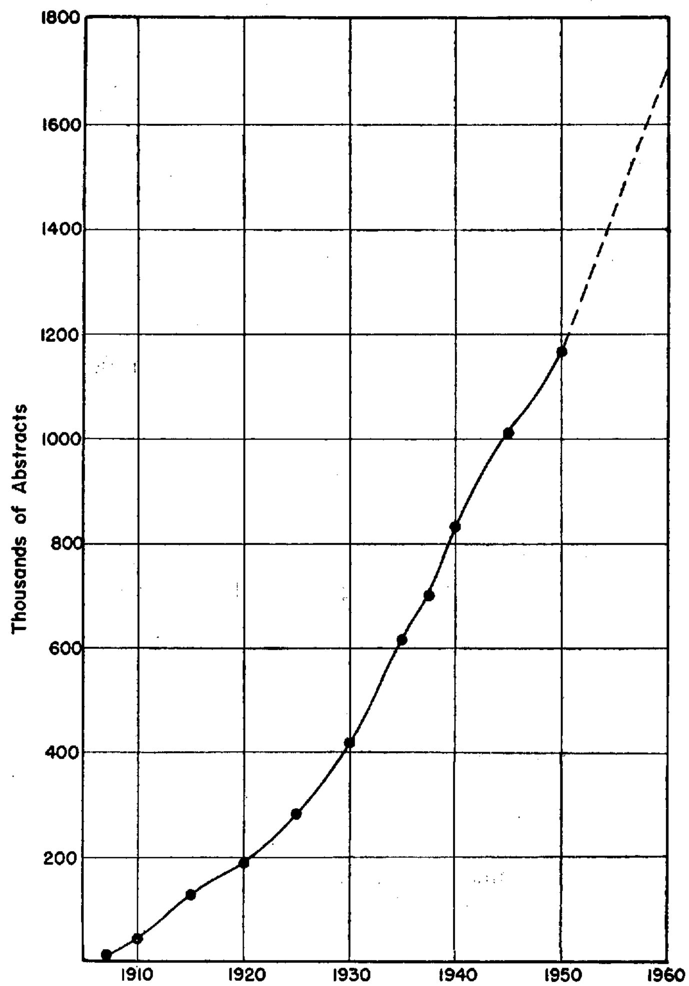

Fig. 1. Cumulative totals of abstracts published by *Chemical Abstracts* (8).

and guiding research and development. Even when information contained in a file is insufficient to provide proof of a relationship between factors, the evidence uncovered by a search to a combination of factors may suffice to aid the imaginative chemist in planning a fruitful line of investigation.

Many of the chemical applications of punched cards have been in the area of the literature of chemical technology, where the information is least organized. The largest segment of this area is the field of private data, the numerous bits of information that constantly accumulate in every individual's and organization's files. This includes research results and other observations, technical production data, private summaries, bibliographics, tabulations and calculations, data for patent applications, complaints and customer reactions, market data, etc. This sort of information is the "raw material" of the literature of chemical technology, and it must be processed, packaged, and distributed before it can be utilized. Other applications of mechanization can facilitate almost every step of the documentary sequence from the time when a chemist first observes a fact until another chemist finds that fact recorded in the literature and uses it, and so starts that autocatalytic sequence all over again (13).

Several types of devices can be used to accomplish sorting and selecting operations that are controlled and determined by independent factors used to analyze information. Of these devices, punched cards have the distinction of being the most widely used at the present time. Impressive savings in time and effort have been achieved with hand-sorted punched cards even in working with files of information relating to a narrow field of specialization, such as the corrosion-resistant properties of nickel alloys (23,24).

Just as hand-sorted punched cards have enabled information files of modest size to be used more effectively, the development of automatic searching equipment provides a basis for making a major advance in the large-scale organization of chemical literature. This possibility comes at an opportune time, for it is becoming clear that the increased volume of chemical research raises grave doubts whether the classical compendia (as Beilstein and Gmelin) will be able in the future to organize chemical knowledge as effectively as in the past (19,20). At the same time, the record-breaking rate of expansion of indexes such as those of *Chemical Abstracts* (as a necessary result of the increase in the number of abstracts published) makes thorough search of the chemical literature an ever more time-consuming task (8). (See Fig. 1.)

Hand-Sorted Punched Cards

How punched cards can be used to search out information relating to any one of several factors (or to combinations of them) is best understood by directing attention to the cards themselves. The most widely used types of hand-sorted punched cards are supplied by the manufacturers with rows of holes along the periphery. Before such cards can be submitted to sorting operations, they must be "punched," that is, the cardboard must be cut away between the edge of the card and appropriately chosen peripheral holes (see Fig. 2). Either a simple hand tool, resembling a conductor's punch, or a power-driven device may be used. Once appropriately punched, the cards may be sorted by using a simple tool resembling a knitting needle mounted in a handle. For sorting, about 200–250 cards are taken, the needle inserted into some hole appropriate to the search, and the cards manipulated so as to fan them out. The cards previously punched at the hole selected are then permitted to fall away from the needle, as shown in Figure 3.

Obviously, such a searching operation has useful significance only if the punching of the hole being searched was assigned some meaning. Once this has been done (see p. 453), a single sorting operation may be used to separate all cards in which any given hole has been punched when the file was prepared for use. The cards which drop on sorting at any one hole may be submitted to further selection by a sorting operation directed to some other hole. Thus, successive sorting operations may be employed to select cards bearing information characterized by a combination of factors which define the search or correlation being undertaken. In place of successive sorting operations with a single needle, several needles may be used simultaneously. To facilitate the use of combinations of needles, manufacturers of punched cards provide specially designed sorting devices (see ref. 5, pp. 45–7, 51).

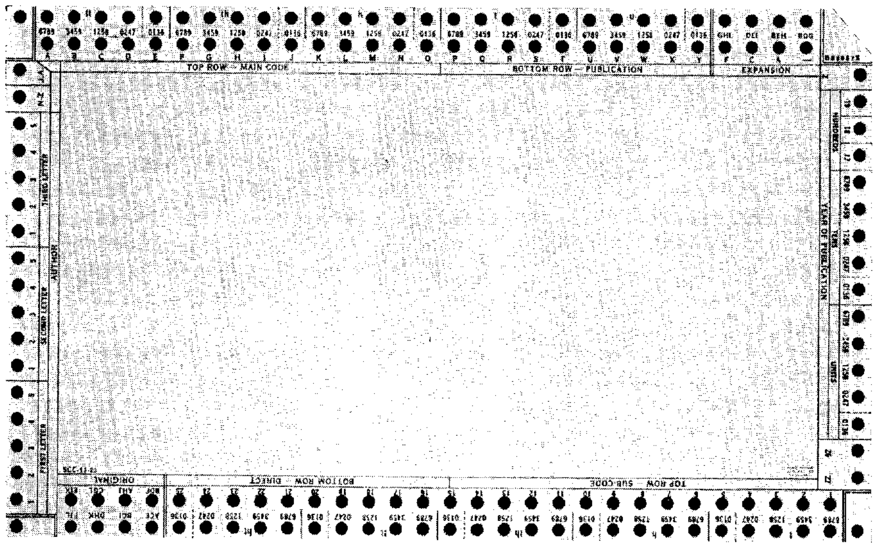

Fig .2. A "Keysort" card specially printed for use in technical abstract files (5).

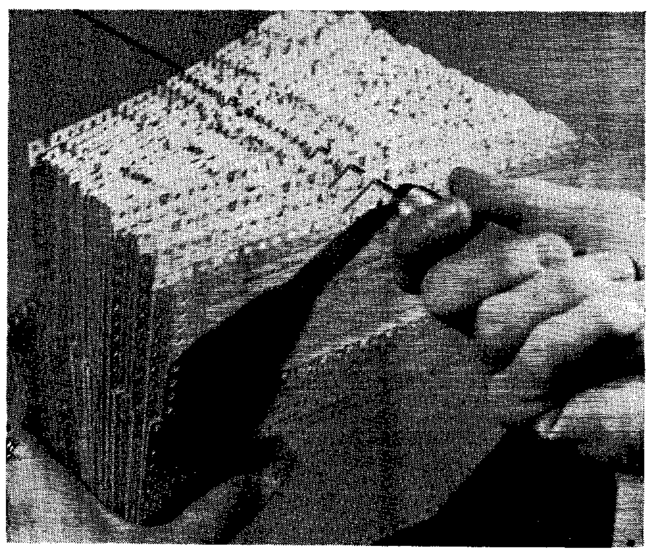

Fig. 3. Hand-sorting a file of punched cards (5).

The holes in edge-notched cards occupy only a small amount of space along the margin, leaving most of the card available for writing or typing references, abstracts, observations, numerical data, etc. (For an example, see Figure 4 under *Microscopy, chemical.*) Moreover, pictures, clippings, and other thin, flat material may be attached to the cards.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150002-0

As already noted, a prerequisite to the use of hand-sorted punched cards is the establishment of a well-defined plan as to how the holes are to be used to indicate the various factors which characterize the information on the card. Such a plan—or *code*, as it is usually called—merits careful thought, as a poorly designed code cannot fail to be a source of dissatisfaction. The simplest approach (sometimes called *direct coding*) is to assign a single hole to indicate a single factor. In setting up such a coding scheme, assignment of factors to holes may be made in any arbitrary manner. However, if neighboring holes are assigned to indicate similar or related factors (for example, different temperature ranges), the punching of the cards is made less prone to error; furthermore, in sorting for a given factor, location of the appropriate hole is facilitated.

It is highly desirable to keep the code system for punched cards as simple as possible. Since direct coding requires a separate hole for each individual factor, the total number of factors available for analyzing information is restricted to the number of holes in the card. This restriction might appear so severe as to make direct coding of little practical value. The ability to direct a search to combinations of factors makes possible surprisingly effective use of properly chosen broad concepts as factors for analyzing information. The secret of developing a simple yet effective punched-card code is the selection, as its basis, of those concepts which (*1*) permit effective analysis of the information to be embraced by the file and (*2*) are effective in defining searching operations to be performed. In selecting concepts for incorporation in the code, attention must be devoted not only to the subject matter to be analyzed but also to the purpose which the punched-card file is to serve. It is not always necessary to code every item of information that is listed on the card.

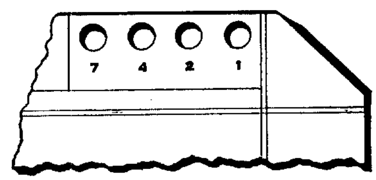

Fig. 4. A numerical sorting field (5).

It is by no means unusual for direct coding to fail to provide for a sufficient number of factors for analysis of the information embraced by a file of hand-sorted punched cards. It then becomes necessary to use somewhat more complex schemes for assigning meaning to the holes. One of these is *combination coding*, in which a combination of two or more holes represents a single factor. The number of combinations that may be set up with the limited fixed number of holes available will obviously exceed the number of individual holes. If each combination is assigned to represent a concept, the number of factors available for analyzing information is increased. However, certain restrictions are encountered as to the use of these concepts, in particular with regard to the number of concepts that can be indicated by the punching on any one card. This point becomes clearer on considering a simple scheme (*7,4,2,1 coding*) for using combinations of holes to indicate the digits 1–9. Any one of the digits 1–9 may be represented by punches in one or two of the four holes shown in Figure 4. Thus, to indicate 1 or 2, we punch the correspondingly labeled hole; for 3, we punch both holes 1 and 2; for 4, the hole so labeled; for 5, the combination 1 and 4; for 6, holes 2 and 4; for 7, the hole so labeled; and so forth. However, only one digit can be punched in each 7,4,2,1 field. For example, punching for both the digits 7 and 2 would be indistinguishable from the punching for 9. If a card were punched for both 8 (1 + 7) and 6 (2 + 4), it would drop no matter which hole was searched with the needle.

By using two fields of four holes each, any one number from zero to 99 may be punched in any one card. With three fields, any number up to 999 may be so punched. One field of four holes is required for each decimal place to code multidigit numbers.

The 7,4,2,1 code was designed for arranging the card file in numerical order. This is done by sorting each hole one after another in the order 1, 2, 4, 7, and placing at the back of the pack the cards which drop after each sort. The sorting is continued from right to left through all the fields in the numerical code.

A punched-card file may be arranged in serial order according to one category ("major item"), and at the same time, under each such item, in serial order according to another category ("minor item"). Thus, a bibliographic file may be arranged in chronological order by date of publication under each major subject covered by the file. Also, the numerical values of any physical or chemical properties of a group of substances can be arranged in order under physiological response, or production output, or any other category. Such multiple-sequence sorting is accomplished by considering each of the major and various minor items as a decimal place in a numerical field, and sequence-sorting as just described. In such a sort, the units are in order under each digit of the tens; the tens are in order under each digit of the hundreds; and so on.

For *selecting* cards coded with a specific number, the 7,4,2,1 code is not the most efficient one. Only those digits requiring two punches can be selected by one searching operation. For example, inserting sorting needles in the 1 and 2 holes in the field shown in Figure 4 will permit only the cards coded 3 to drop. However, if the 2 hole is sorted with a single needle, the cards which drop are not only the ones coded 2 but also those coded 3 (2 + 1), 6 (2 + 4), and 9 (2 + 7). These unwanted cards can, of course, be eliminated by submitting the cards that previously dropped to supplementary searching operations directed to holes 1, 4, and 7. In these supplementary searches, the unwanted cards drop out and are eliminated.

This need for supplementary searching operations can be eliminated by adding to the field one more hole which may be labeled SF (signifying "single figure") and which is punched along with the 7, 4, 2, or 1 when one of these four digits is to be coded. Then, two sorting needles in each field will select in a single sorting operation just the cards coded for the number being sought. Thus, for the digit 6, needles would be inserted in the 2 and 4 holes, while for the digit 2, needles would be inserted in the 2 and SF holes. For serial sorting as described above, the SF hole is simply ignored. Zero may be coded by punching 7 and 4. When serially sorted, these zero cards will follow those coded 9. They may be moved manually to their proper position preceding 1.

Another variation of this general type of coding, with one entry in a given field, provides the possibility of coding any one of the 26 letters of the alphabet in a group of eight holes, which are labeled A,B,D,G,K,P,V, and SF. The punching scheme is shown in Table I. Unused combinations (V + K and V + P) may be assigned special meaning. Thus, V + K might be punched to indicate anonymous authorship. Other variations of this general punching method involve the use of holes arranged in a double row to represent digits or letters (7).

A related, though distinctly different, type of coding is also based on assigning a combination of holes to indicate a single factor. This method in its simplest variation uses one rather large field in which each hole is designated by some convenient symbol, for example, a letter of the alphabet for each of 26 holes set up as a single field. In such a field, three-letter combinations might be punched to designate various animals, for example, CAT for cat, DOG for dog, COW for cow, HOR for horse, MON for monkey. Let us take as an example a card bearing information involving these five ani-

TABLE I. Coding the Alphabet in a Group of 8 Holes.

Letter	Holes punched	Letter	Holes punched
A..............	A + SF	N..............	K + D
B..............	B + SF	O..............	K + G
C..............	A + B	P..............	P + SF
D..............	D + SF	Q..............	P + A
E..............	D + A	R..............	P + B
F..............	D + B	S..............	P + D
G..............	G + SF	T..............	P + G
H..............	G + A	U..............	P + K
I..............	G + B	V..............	V + SF
J..............	G + D	W..............	V + A
K..............	K + SF	X..............	V + B
L..............	K + A	Y..............	V + D
M..............	K + B	Z..............	V + G

mals and no others. In the proper field in this card, the three-letter designations corresponding to the five animals would be punched. Obviously, if a search were directed to any of these three-letter combinations, the card would drop and thus be selected. This punching scheme is often called *superimposed coding*, since several combinations of holes are punched in a single field.

The principal advantage of superimposed coding is the possibility of using a very large number of concepts as a vocabulary for analyzing information. There are limits, however, to the number of entries that can be punched in any one card. In the example above, the card would respond to the letter combination CAN, composed of the CA from CAT and the N from MON. This means that the example card would have been selected by a search to CAN for "canary," even though the card makes no mention of that bird. The appearance of such "unwanted" or "extra" cards can become very troublesome if superimposed coding is used carelessly. Mathematical analysis, supplemented by experience, has provided a basis for estimating the chances of undesirable results and for avoiding them by care in designing and using superimposed coding (25). Even with care in designing the code, it is not advisable to conduct superimposed coding so as to punch many more than one-third of the holes in a field. In our example card, 11 holes out of 26 were punched, and such a card may be expected to appear occasionally as an unwanted card. If more holes had been punched, this undesirable possibility would have been increased.

As might be expected, a number of variations of superimposed coding have been developed. One of these employs four fields of 26 holes each and designates entries by four-letter combinations, such as SUGA for sugar, RESI for resin. One field is used for the first letter in any four-letter code designation, another field for the second letter and so also for the third and fourth letters. This coding scheme permits as many as nine or ten different entries to be made per card without causing undue probability of extra cards in excessive numbers (25).

Various combinations of the three basic coding methods outlined above have been used with different punched-card files. For example, one field using direct coding may be used to indicate general categories, to which search may be directed frequently, while another field set up to use superimposed coding may be punched for more specific terms used to indicate important details. In addition, numeric or alphabetic coding may be employed to provide such data as year of publication, or author's initials.

Hand-sorted punched cards offer considerable scope for ingenuity in setting up schemes for analysis of information.

The emphasis given so far to the mechanics of using hand-sorted punched cards, though unavoidable, incurs the risk that the importance of related intellectual problems may be underestimated. The use of the cards as tools for searching and correlating information presupposes the analysis of this information in terms which provide a basis both for characterizing the subject matter and for conducting the mechanical searching operations. Here a word of warning is in order.

The beginner must be particularly careful to keep his coding scheme as simple as possible. He must not permit enthusiasm for the flexibility of the punched-card method and its possibilities to mislead him into overdoing his efforts to exploit them. It is particularly important to remember that mechanical sorting operations cannot do more than select certain items, which must then be examined personally to determine their significance. If, as a result of mechanical sorting, cards relating to items of borderline interest are also selected, no great disadvantage is incurred as long as the number of such cards is not excessive. In practice, this means that coding for hand-sorted punched-card files need not be designed to select information with pin-point accuracy. If excessively fine analysis and coding were attempted, these operations might become so time-consuming that much of the advantage of using punched cards would be lost.

Although precise rules for developing a punched-card code cannot be given, certain guiding considerations can be cited. One important principle is to define with care the purpose or purposes which the file is to serve. If uncertainty exists in this connection, it may be advisable to avoid all punching at first and to use the unpunched cards for a time as ordinary file cards. Or perhaps certain concepts will be so obviously important from the start that they should be coded and punched while judgment is reserved as to how the finer details of the system are to be worked out. Another important consideration is the anticipated size of the file, that is, the number of cards it will probably contain eventually. For a large file, it may be well to employ more detailed coding in order that mechanical searching may eliminate more of the cards bearing information of borderline interest. Uncertainty as to the eventual size of the file is an argument in favor of holding some of the coding possibilities in reserve. A supporting argument may be the likelihood of new trends requiring the coding of new concepts in the future.

In thinking through the various considerations involved in setting up a code for hand-sorted punched cards, it is particularly helpful to study the methods successfully developed by others to meet their various information problems (see ref. 5, pts. II and V).

The advantages of hand-sorted punched cards may be summarized as follows: Once appropriately punched, they can be kept in any order in a file from which can be selected all cards bearing a reference to any one of the factors included in the code, or to any combination of such factors. Only one card is required for each reference. If the coding includes authors' initials, date of publication, or serial number (as of a patent), then arrangement of the cards in alphabetical or numerical order in accord with these categories is readily accomplished by simple mechanical operations. Finally, hand-sorted punched cards and the tools used to punch and to sort them are low in cost (26).

Systems Related to Hand-Sorted Punched Cards

The "Flexisort" system employs the same coding and sorting techniques already discussed. This system, however, does not use specially prepared cards with rows of marginal holes perforated by the manufacturer. Instead, a special punch is employed

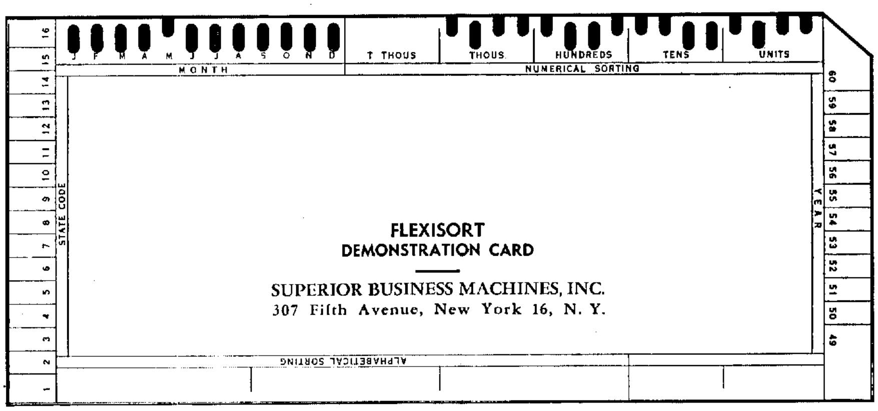

Fig. 5. A punched card employing the "Flexisort" system, showing the arrangement of holes and notches (5).

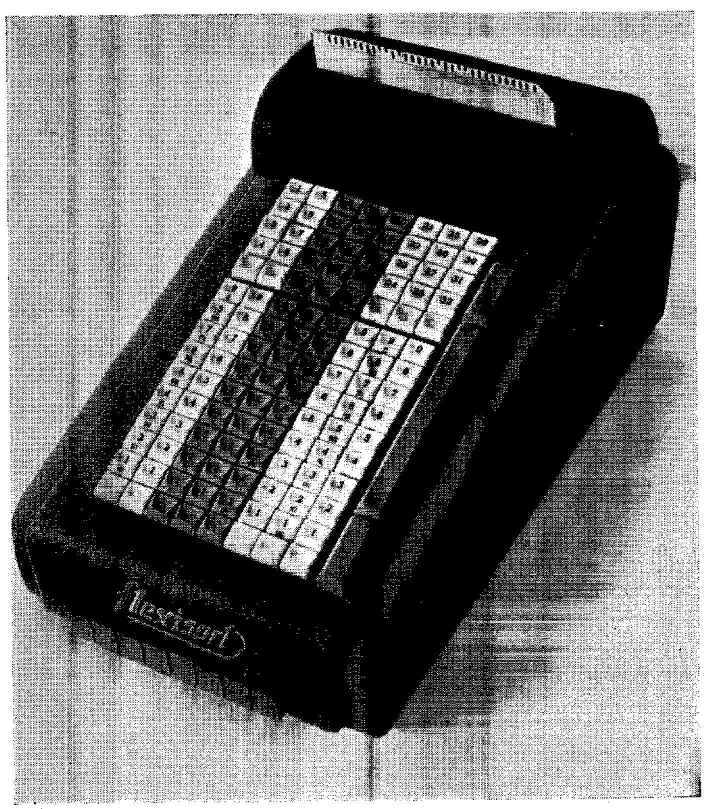

Fig. 6. "Flexisort" punch (5).

to cut, simultaneously, marginal holes (corresponding to unpunched holes in prepared cards) and notches (corresponding to punched holes). (See Figs. 5 and 6.) The keyboard of the punch includes alphabetical characters as well as numerals. Perforating and punching one margin of the card involves depressing the keys corresponding to the desired letters and numerals, and then depressing the motor bar. The machine can perforate and punch cards of any size or weight of card stock. This system may also be used to convert existing files of ordinary cards into marginal-hole punched-card files (see ref. 5, pp. 52–4).

The "Zatocoding" system resembles manually sorted punched cards in that slots are cut to the periphery to render the cards responsive to sorting. The Zatocoding cards differ in that they are supplied by the maker without holes perforated at the slotting positions. As a consequence, these cards cannot be sorted with a simple hand tool as can "Keysort," "E-Z Sort," "Pathfinder," and other widely used, edge-notched punched cards suitable for manual sorting, but a vibratory sorting machine is used instead. In preparing this machine for making a search, rods of small diameter are inserted in positions which correspond to slots cut in the cards to indicate significant factors. The Zatocoding system employs randomly selected numbers as a basis for superimposed coding.

Automatic Machines and Literature Searching

While hand-sorted punched cards are well suited for files of moderate size, the amount of manual effort required to use a large file may become burdensome. With files containing about 10,000 cards or more, the possibility of using automatic equipment will inevitably suggest itself. Large files may present opportunities for automatic machines to effect important savings in clerical time.

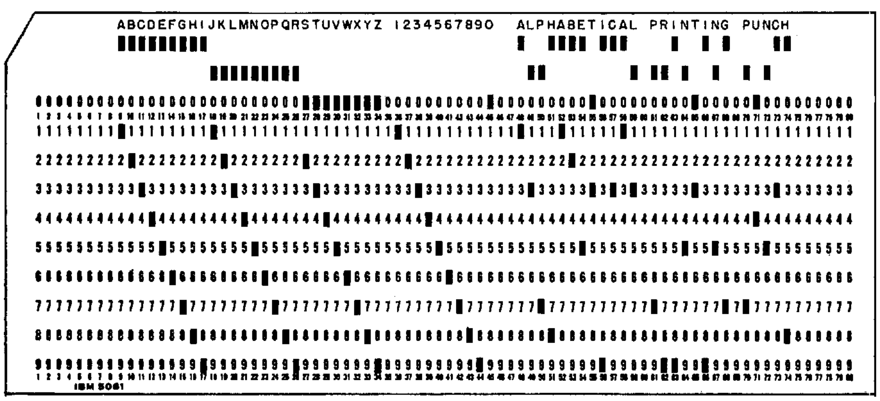

Fig. 7. Standard IBM card (5).

The manual effort required for handling large masses of cards is not the only consideration which may make automatic equipment appear more attractive. A large file covering a broad field may prove difficult to manage because of the limited number of coding possibilities offered by hand-sorted cards. Machine-sorted cards work with a larger number of holes and consequently it is logical to look in that direction when establishing a system to embrace a broad field, such as inorganic chemistry.

A standard punched card for accounting purposes is shown in Figure 7. On this card the digits 0–9 are printed in 80 vertical columns and the column number itself is printed immediately below the 0 and 9. Above the 0 in each column there are two additional positions—sometimes called the X and Y positions—which may also be punched. The holes in the "IBM" and "Remington Rand" cards permit, respectively, electrical and mechanical contacts to be made which actuate the mechanisms that perform the desired operations when the cards are fed automatically through the machines. For example, all cards perforated in a given manner can be selected from the file. Also, the whole file may be sorted into numerical or alphabetical sequence. The symbols coded on the card can be printed on the same card with each symbol above the column in which it is coded, or on another sheet of paper, and numerous other operations performed. References, abstracts, and other data may be entered in the ⅛ inch of clear space between the rows of perforations in machine-sorted cards by use of a micro-typewriter. In some applications, a number of columns at one end of the card are left unpunched, and written or typed data are entered in that space.

However, it must be noted at this point that nearly all the automatic punched-card machines available commercially at present (1951–1952) have been developed primarily for the purpose of conducting accounting operations, for example, inventory control, payroll preparation, or compilation of business statistics (18,26). As a consequence, such equipment suffers from certain limitations which are not inherent in automatic searching machines in general or punched-card techniques in particular. This point is of sufficient importance to warrant more thorough discussion.

Punched-card machines for accounting are so designed that searching and selecting operations are directed to a hole (or holes) in a specific column or—with certain special machines—to certain combinations of columns. In order that card-operated accounting machines can search effectively, the column or columns in which a given entry has been punched must be known. This means that the columns in which individual criteria or factors are to be punched must be planned in advance. A column or group of columns must be assigned to each criterion or group of criteria. There is the further restriction that only one member of a group of criteria may be punched in the portion of the card reserved for that group. This latter limitation is not particularly troublesome for accounting purposes, as a typical sale of some commodity involves some single unit price (to be punched in a field of columns reserved for price), a certain single amount of the commodity, a certain single purchaser, and other single-valued pieces of information, each of which can be punched into appropriately assigned fields.

Punched cards designed for accounting have been applied to files relating to a relatively narrow field of specialization, such as infrared absorption spectra, without the restrictions inherent in the assignment of fixed fields becoming excessively troublesome (15). In such cases, appropriate fields may be set up for the different types of data to be indicated by punching. Furthermore, there is a high degree of probability that entries will be made in a majority of the fields of any one card. The situation becomes increasingly less favorable as the range of subject matter is broadened, with inevitable extension in the range of criteria that are needed for analysis of information and must be accommodated on the card. Finally, as studies at the U.S. Patent Office have shown, for collections of information extending over a very broad field, the limitations characteristic of standard accounting machines are such that their use is of very doubtful practicality (2). These limitations, it should be emphasized, apply to standard accounting machines and are not necessarily inherent in all automatic punched-card machines.

On the occasion of the Diamond Jubilee Meeting of the American Chemical Society in September, 1951, the International Business Machines Corporation announced that they had constructed experimental models of punched-card machines especially designed for searching and correlating information (6). A newly developed coding method eliminated the necessity of working with fixed fields. Instead, any one column or sequence of columns could be punched to indicate a code designation which might be any one of various numerals or letters. The punching was set up so that a predetermined combination of five holes in any column always indicated the same letter, or numeral, as the case might be. It thus became possible to spell out any word—or, more generally, any meaningful sequence of coding symbols—by punching them in successive columns of a card. This type of punching might be compared with the patterns of raised dots used in Braille. Searching is accomplished by reading the entire card photoelectrically and detecting those patterns of holes which spell out the criteria that characterize the information being sought. A plurality of photocells permits searching to be directed to combinations of criteria. The new searching machine, in addition to identifying wanted criteria, also permits their relationship to each other to be specified. Thus, the machine may be directed to accept all cards punched for any one of several criteria. Such a search involving six criteria might be symbolized as A + B + C + D + E + F. Alternatively, it might be specified that only cards bearing all six criteria would be accepted and such a search might be represented by A.B.C.D.E.F. Combinations may be specified such as (A+B).(C+D+E).F. Another possibility is to require that some criterion—or one or more of a group of criteria—shall be absent. Examples of such searches might be (A+B).(C+D+E)−F or (A.B+C.D)−(E+F).

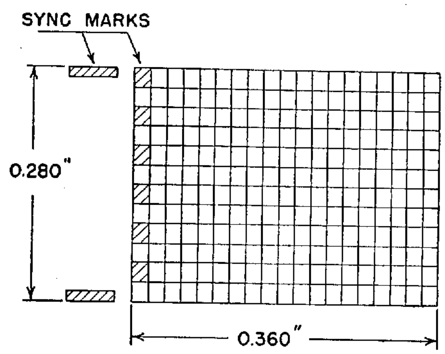

Fig. 8. Layout of coding area for rapid selector (25a).

Previous discussion has been concerned with punched cards as their applications have constituted the principal advance made up to the present time in the mechanization of literature searching. The importance of making information available when needed has attracted the attention of inventors in the field of electronics, and several ingenious machines have been either constructed or proposed. Among such machines, the rapid selector has attracted much attention (9,10,22).

In order to understand the principle on which the rapid selector operates, it is necessary to visualize, first of all, a reel of motion-picture film in which an imaginary line running longitudinally down the center of the film divides the successive frames into halves. On one side of the line, microphotographs of abstracts are entered one after another. The coding which indicates the criteria characterizing the subject matter of any one abstract is entered on the other side of the line in a frame displaced nine frames rearward from the frame bearing the corresponding abstract. The reason for this displacement depends on the detailed mechanics of the searching operation. Before describing the latter, it should be noted that the coding which indicates the subject matter of an abstract consists of a block of seven digit numbers entered one after another in the appropriate half frame of the film. Each half frame used as a coding area is laid out in small squares arranged in twelve rows of eighteen squares each.

Each number is entered as a pattern of opaque and transparent squares in two successive rows of holes, as shown in Figure 8. Each coding area is thus able to accommodate six coded entries (criteria). If more entries are needed for any one abstract, it may be repeated on the film as many times as necessary and the additional entries made in corresponding coding areas.

As set up recently for tests at the Department of Agriculture in Washington, the rapid selector permitted searching for any one desired entry as follows. First, the seven-digit number corresponding to the desired entry was determined by consulting the code book. This code number was represented, of course, by a certain pattern of opaque and transparent squares. A searching mask was prepared in which holes were punched to correspond to the opaque squares in the pattern to be detected. The reel of film was then run through the machine and light passed through both the mask and the code areas of the moving film until such time as the punching in the mask corresponded exactly to the number being sought. The resulting momentary blackout, detected with the aid of one or more photocells, activated an electronic circuit and then, with the aid of high-speed flash photography, a picture of the abstract to which the coded entry referred was taken. This picture was recorded on a separate reel of film that, on subsequent development, provided the result of the search.

This arrangement, though capable of operating at a rate of 20,000 frames per minute, permitted only one searching mask to be used. Furthermore, the film produced as the result of a search could not be run through the machine again. As a consequence, the machine as first set up at the Department of Agriculture did not permit searches to be directed to a combination of entries. This serious deficiency could be overcome in part by applying the principle of superimposed coding to the rapid selector. The most effective way to do this would be to redesign the scanning element so that the search mask would extend to cover all the squares used for coding. Methods developed for using superimposed coding with hand-sorted punched cards could then be readily applied to the rapid selector (25a). In this way, it would be possible to conduct searches directed to combinations of criteria defined symbolically by A.B.C.D.E.F. However, more extensive redesign would be necessary to conduct searches involving alternate criteria, symbolized by A+B+C+D+E+F, or specification of the absence of one or more criteria, for example (A.B+C.D)−(E+F). A multiplicity of photocells to detect individual criteria would be needed as well as supplementary electronic circuits to detect the specified relationships between the photoelectrically identified criteria.

The newly developed IBM equipment and the rapid selector operate on the principle of matching patterns, either of holes in cardboard or of spots (transparent and opaque) in film. During recent years, patterns of magnetized spots on tapes have been used for both input and output of data with electronic digital computers. These machines are not only capable of performing arithmetical operations at very high speeds, but they can also follow through elaborate programs which the machines may themselves alter in a predetermined fashion, depending on the results obtained in carrying out successive portions of a routine. These machines are also provided with so-called internal memory units, which enable intermediate results to be stored temporarily pending further processing. Electronic digital computers owe their usefulness to the very high speed at which various individual operations are performed. The digital computer known as Whirlwind I requires only 60 microseconds for each operation (1).

Electronic digital computers can be used to search an information file provided its component items have been analyzed as to appropriate criteria and the latter encoded as patterns of magnetized spots on the input tapes. When searching, the machine directs attention to successive blocks of entries pertaining to separate items in the file. In the examination of any one block of entries, each of the latter must be checked one by one for identity with the criteria used to specify the scope of the search. When an identity is found, that fact must be recorded in the internal memory, pending the result of search for other criteria. Finally, if all the criteria are identified and if the relationship between them is found to be as specified, the serial number of the item is typed out by the machine's output unit. Since digital computers perform individual identifying operations one by one, the programming required is complicated and the time consumed in searching the block of entries pertaining to a single item of information is several times longer than that required for photoelectric scanning of the same entries punched in a card or entered on photographic film.

The explanation for the surprising slowness of digital computers when used as searching machines is to be found in the fact that the individual steps in the unavoidably lengthy program must be performed one after another. Component units and parts of a digital computer could be used to construct a searching machine in which each encoded entry as it was read from the input tape would be simultaneously inspected by a bank of comparator units, each of which would be set for a different criterion used to define the scope of the search. Identification of an encoded entry from the tape with the specified search criterion would cause the comparator unit to transmit a pulse to another section of the machine. Here pulses from the comparator units would be caused to interact to determine whether the entries identified were such as to satisfy the search requirements with regard to relationships between criteria. Since the comparator units operate in parallel, no elaborate programming would be required and very high searching speeds could be attained. It is estimated that five million blocks of encoded entries could be searched per hour (1).

This summary of some of the more important characteristics of automatic searching equipment may arouse a variety of questions in the mind of the thoughtful reader. What practical use could be made of such machines? What circumstances would determine which type of machine should be used? What procedure would prove best suited for analysis of information preparatory to search by automatic machines? Experience in using such machines is as yet too limited to provide more than partial answers to these questions.

With regard to practical use of such machines, it must be noted, first of all, that a considerable investment in human effort would be required to analyze information and encode the entries selected as appropriate for characterizing it. Such an investment would be most profitable for organizations in which a large file of information is repeatedly searched from different points of view. An example of such a situation is furnished by the Patent Office, where the very large number of issued patents constitute a file which must be searched from different points of view in examining each year tens of thousands of applications as to their patentability. Successful experiments with machine searching at the Patent Office have provided a basis for impressive savings in time (3).

Scientific and technical publications in the realm of chemistry also constitute a very large file which is growing at a rapid rate, as shown by Figure 1. Searching this file is an activity which is essential to the planning and accomplishment of chemical research and development. Hundreds of thousands of man-hours are devoted each year by chemists to the routine operations involved in literature searching. Distaste for the drudgery of literature searching can be even more costly in terms of misguided or even useless experimental work. Recently developed machines could be used to relieve this drudgery, and realization of this fact has been the basis for much activity which up to the present has been largely in the planning stage (17,19).

Large-scale utilization of automatic machines for searching chemical literature cannot be undertaken lightly. Although eventual savings in human effort—and hence in money—would be large, an investment of considerable size would be required for preparatory work. In particular, the papers, patents, reports, etc. to be searched would have to be analyzed and encoded. Both the magnitude of the expense and the effectiveness of machine searching operations would be determined to a large degree by the system used for the analysis and encoding. Establishing the most effective system is a complex problem which is still under investigation (18,19). A few of the principles guiding this development will be summarized briefly.

One basic consideration is that the system must be capable of being used conveniently and effectively by human beings. Overlooking this very simple consideration might lead to developing a system which, however attractive in theory, would prove worse than useless in practice. This consideration alone suffices to suggest that the system for analyzing information preparatory to machine searching should be as nearly similar to indexing as possible. Such an approach would permit the person analyzing papers, patents, etc. to use the terminology either found in the documents or suggested by them as a basis for indicating their content. There would be no need to attempt the difficult feat of keeping in mind a highly ramified coding system and applying it in detail to papers, patents, etc. being scrutinized for analysis.

Another consideration is the fact that the automatic machines, although able to identify properly encoded entries in cards, film, or magnetic tape, are incapable, except in a very limited sense, of interpreting the entries. Thus, if Massachusetts were designated for machine searching by MASS, this symbolism of itself would not permit direct identification in a search directed to New England or to the U.S.A. Such direct identification does become possible if appropriate symbolism is chosen. Part of such a system is as follows:

Entry	Symbol
United States of America	US
New England	USNE
Massachusetts	USNEMA
Connecticut	USNECO
Vermont	USNEVT
Maine	USNEME
New Hampshire	USNENH
Rhode Island	USNERI
Southern	USSO
North Carolina	USSONC
Alabama	USSOAL

With this approach, indexing for Massachusetts is followed by encoding as USNEMA, thus making the code for New England, USNE, and for the United States of America, US, available for direct searching. Generalizing, it may be said that the encoding operation may be used as a means of providing generic terms for machine searching while permitting the analysis of documents to be accomplished by citing terms either found in the document or immediately suggested by its subject matter (16).

This method of encoding has obvious limitations. Coding Massachusetts as USNEMA does not indicate that it is an industrial state. It would, of course, be possible to set up the coding to do this, but clearly some restrictions must be imposed if the code designations are not to become so lengthy as to be unwieldy. The considerations involved in selection of the most appropriate basis for arraying terminology in setting up a coding system would seem to be rendered more complex by the fact that the relationship between different types of terms, such as names of substances, processes, or attributes, may provide a useful distinction to incorporate in the code. Thus, various symbols might be affixed to a basic term such as "polymer" to indicate: (*1*) a polymerized substance, (*2*) the process of polymerization, or (*3*) the abstract idea "polymeric." While such special symbolism might, at first glance, seem likely to complicate the code, actually it can lead to simplification if it is used with discretion.

The methods outlined above, although making possible searches defined by terminology of varying degree of generality or specificity, are not sufficient to remove certain ambiguities from the searching operations. For example, the mere encoding of chemical substances would not suffice to distinguish between reaction products and starting materials, nor between donor and acceptor, for these have the type of relationship sometimes described as asymmetric. Such ambiguity may be resolved either by attaching meaning to the order in which encoded entries are made or by employing special symbols to distinguish reaction products from starting materials, donors from acceptors, etc. The latter method is simpler and more reliable with equipment available for experimentation at present.

The **encoding of chemical structural formulas** is a particularly knotty problem. Two somewhat different methods have been used to attack it. One method is to cite the component structural units, for example, benzene rings, hydroxyl groups, or chlorine atoms, together with appropriate symbols to show how the groups fit together. The other method starts at some one point in the molecule and cites successive atoms encountered on proceeding from one atom to the next, by means usually of letters, numerals, and certain conventional symbols (such as punctuation marks).

A commission of the International Union of Pure and Applied Chemistry was appointed in 1946 to study the problem of coding structural formulas. Nine different coding and classification systems (27,29,30–34,36–38) came to this commission's attention, of which the Dyson cipher (31) was judged to be most promising (35) (see also *Nomenclature*). At the present time, a committee of the National Research Council is investigating possibilities of improving the Dyson system (28). It is hoped that international agreement as to a generally acceptable code may be achieved in the near future.

Recording and Reproducing Techniques

Although the sorting and selecting operations effected by punched cards and similar devices save time and drudgery by directing attention to needed items of information, they are not of themselves sufficient to provide needed information. With hand-sorted punched cards, for example, information must be typed or otherwise entered on the cards. Typing abstracts, data, or the like on a large number of cards may prove excessively time-consuming and expensive. A simple alternate method is to seal data sheets, clippings, or the like to the cards with the aid of appropriate adhesive.

Kodak dry-mounting tissue, a thermoplastic sheet waxed on both sides, has been used with excellent results. After sandwiching the tissue between the card and data sheet, clipping, or the like, sealing is effected by hot pressing. This method was used in setting up the previously mentioned punched-card file on the corrosion-resistant properties of nickel alloys (23,24).

Photography in one variation or another may be helpful in copying information onto punched cards (43). (See also *Photoreproduction.*) The diazo photographic processes "Ozalid" and "Rétocé" warrant special mention in this connection. The electrostatic electrophotographic method known as "Xerography" may also prove useful. With this method, the information would be transferred onto the xerographic plate and from there directly to the punched card. Facsimile reproducers and the "Thermofax" process might also find application in preparing punched-card files if appropriately treated cards were available.

Transcription processes may also play an important role in using hand-sorted punched-card files. Once sorting operations have selected out certain cards, copying off the information may be desirable. The need to prepare a duplicate file of hand-sorted cards may also arise and this may require transfer of information entered on the original file of cards (43).

Microphotography may also be used to advantage in connection with mechanized literature searching. With punched cards, microfilm inserts permit the complete text of a paper of moderate length to be entered on a single card. Drawings, photographs, and other pictorial material may also be entered as microfilm inserts (see ref. 5, pp. 71–74).

With large files of information, it may prove advisable to use automatic equipment for searching and microphotography for maintaining information files. This would permit an advantageous division of labor. The high-speed searching device would identify pertinent documents while storage of the full text of documents for reference might be accomplished in photographic form (microfilm or "Microcards") (40–42,44,45).

Locating previously identified items in an extensive microphotographic file might eventually become burdensome. It should be possible to develop auxiliary equipment which would be activated by serial numbers of desired documents in such a way as to produce copies of the desired documents. Up to the present, the need for such a machine has not become acute, but its construction should not present insuperable problems.

In addition to providing a means of convenient storage of documents, microphotography may facilitate distribution of the results of machine searching to interested persons. Here the pioneer work of Kuipers suggests interesting possibilities (39).

Modern techniques for storing information and for searching, by providing access to needed facts and data, are powerful adjuncts to human memory. In addition to bringing needed information to notice, the establishment of correlations, the evaluation of data, and the planning and guiding of research may be facilitated (21).

Hand-sorted punched cards are supplied by: The McBee Co., Athens, Ohio ("Keysort"); E-Z Sort Systems, Ltd., 45 Second St., San Francisco, Calif. ("E-Z Sort"); and Charles R. Hadley Co., 330 North Los Angeles St., Los Angeles, Calif. ("Pathfinder"). The Zator Co., 79 Milk St., Boston, Mass., supplies Zatocoding cards.

Machine-sorted punched cards and equipment for use with them are manufactured by International Business Machines Corp., 590 Madison Ave., New York, N.Y., and by Remington Rand, Inc., 315 Fourth Ave., New York, N.Y.

Various devices for preparing and using cards with film inserts are available at Film 'N File, Inc., 330 West 42nd St., New York, N.Y.

Bibliography for Mechanized Searching

(1) Bagley, P. R., and Perry, J. W., "Applicability of Newer Electronic Techniques to Information Searching," *Abstracts of Papers, Diamond Jubilee (120th) Meeting, Am. Chem. Soc. (New York)*, Sept. 1951, pp. 2F–3F.

(2) Bailey, M. F., and Cochran, S. W., "Patent Searching—General Files," in Casey and Perry (see ref. 5), pp. 367–77.

(3) Bailey, M. F., Lanham, B. E., and Leibowitz, J., "Mechanized Searching in the U.S. Patent Office," *Abstracts of Papers, Diamond Jubilee (120th) Meeting, Am. Chem. Soc. (New York)*, Sept. 1951, p. 2F.

(4) Breger, I. A., "Applications of Simple Coding Procedures to a Specific Problem," in Casey and Perry (see ref. 5), pp. 27–37.

(5) Casey, R. S., and Perry, J. W., *Punched Cards. Their Applications to Science and Industry,* Reinhold, N.Y., 1951. Figures 2–7 courtesy Reinhold Publishing Co.

(6) *Chem. Eng. News*, **29,** 4214 (1951).

(7) Cox, G. J., Bailey, C. F., and Casey, R. S., *J. Chem. Education*, **24,** 65–70 (1947).

(8) Crane, E. J., *Chem. Eng. News*, **23,** 1757 (1945); **25,** 1188 (1947); **26,** 2190 (1948).

(9) Engstrom, H. T., "Microfilm Selection Equipment in Information Work," *Ind. Eng. Chem.*, **42,** 1460–61 (1950).

(10) Engstrom, H. T., *Report for the Microfilm Rapid Selector,* No. **97313,** Eng. Research Associates, St. Paul, Minn., and Arlington, Va.

(11) Forrester, J. W., "High-Speed Electronic Computing Devices," *Ind. Eng. Chem.*, **42,** 1461 (1950).

(12) Frear, D. E. H., "Correlation of Research Data and Establishment of Cause and Effect Relationships," in Casey and Perry (see ref. 5), pp. 305–10.

(13) Hill, N. C., Casey, R. S., and Perry, J. W., *Chem. Eng. News*, **25,** 970 (1947).

(14) Isbell, A. F., "An Improved Punched-Card System for Handling Scientific Information," *Abstracts of Papers, Diamond Jubilee (120th) Meeting, Am. Chem. Soc. (New York)*, Sept. 1951, p. 3F.

(15) Keuntzel, L. E., *Codes and Instructions for Wyandotte Punched Cards Indexing Infrared Absorption Spectrograms,* Wyandotte Chemicals Corp., Wyandotte, Mich., 1951.

(16) Perry, J. W., *Am. Documentation,* **1,** 133–39 (1950).

(17) Perry, J. W., *Chem. Eng. News,* **26,** 3118–19 (1948); **27,** 754–56 (1949); **28,** 3789 (1950); **29,** 4498 (1951).

(18) Perry, J. W., "Scientific Aids for Literature Searching," *Chem. Eng. News,* **30,** 158 (1952).

(19) Pietsch, E. H. E., "Future Possibilities of Applying Mechanized Methods to Scientific and Technical Literature," in Casey and Perry (see ref. 5), pp. 427–55.

(20) Pietsch, E. H. E., *Nachrichten für Dokumentation,* **2,** 38–44 (1951).

(21) Ranganathan, S. R., and Perry, J. W., *J. Documentation,* **7,** 10–14 (1951).

(22) Shaw, R. R., *J. Documentation,* **5,** 164–71 (1949).

(23) Voigt, L. E., *Corrosion,* **4,** No. 12, 582–89 (1948).

(24) Voigt, L. E., and Rea, A. E., "Activating Unwieldy Files," in Casey and Perry (see ref. 5), pp. 79–90.

(25) Wise, C. S., "Mathematical Analysis of Coding Systems," in Casey and Perry (see ref. 5), pp. 276–302.

(25a) Wise, C. S., and Perry, J. W., *Am. Documentation,* **1,** 76–83 (1950).

(26) Zeisig, H. C., Jr., and Martin, P. T., "Commercially Available Punched-Card Systems, Equipment and Supplies," in Casey and Perry (see ref. 5), pp. 39–75.

CODING OF CHEMICAL STRUCTURAL FORMULAS

(27) *A Method of Coding Chemicals for Correlation and Classification,* Chemical-Biological Coordination Center, Natl. Research Council, Washington, D.C., 1950.

(28) Berry, M. M., and Perry, J. W., "Notational Systems for Structural Formulas," *Chem. Eng. News,* **30,** 407–10 (1952).

(29) Buhle, E. L., Hartnell, E. D., Moore, A. M., Wiselogle, L. R., and Wiselogle, F. Y., *J. Chem. Education*, **23**, 375–91 (1946).
(30) Cockburn, J. G., *The Newcastle System of Representation of the Formulae of Organic Compounds* (private communication).
(31) Dyson, G. M., *A New Notation and Enumerating System for Organic Compounds*, 2nd ed., Longmans, Green, N.Y., 1949.
(32) Frear, D. E. H., "Comprehensive Coding Schemes for Chemical Compounds," in Casey and Perry (see ref. 5), pp. 311–28.
(33) Gordon, M., Kendall, C. E., and Davison, W. H. T., "Chemical Ciphering: A Universal Code as an Aid to Chemical Systematics," *J. Proc. Roy. Inst. Chem. G. Brit. Ireland*, **1948**, p. 46.
(34) Gruber, W., "Die Genfer Nomenklatur in Chiffren und Vorschläge für ihre Erweiterung auf Ringverbindungen," *Angew. Chem., Beihefte*, **58**, 72 pp. (1950).
(35) Patterson, A. M., *Chem. Eng. News*, **29**, 4078, 4116 (1951).
(36) Silk, J. A., *A New System of Organic Notation* (private communication).
(37) Weerden, W. J. van, *Classification of Organic Compounds* (private communication).
(38) Wiswesser, W. J., *Pictorially Direct Chemical Structure Symbols* (private communication).

RECORDING AND REPRODUCING TECHNIQUES

(39) Kuipers, J. W., "Microcards and Microfilm for a Central Reference File," *Ind. Eng. Chem.*, **42**, 1463–67 (1950).
(40) Power, E. B., *Am. Documentation*, **2**, 33–39 (1951).
(41) Rider, F., "Microcards, a New Form of Publication," *Ind. Eng. Chem.*, **42**, 1462 (1950).
(42) Rider, F., *Am. Documentation*, **2**, 39–44 (1951).
(43) Stanford, S. C., and Gull, C. D., "Transcription Problems in Preparing and Using Punched-Card Files," in Casey and Perry (see ref. 5), pp. 395–401.
(44) Tate, V. D., *Am. Documentation*, **1**, 91–99 (1950).
(45) Tauber, M. F., "Problems in the Use of Microfilms, Microprint, and Microcards in Research Libraries," *Ind. Eng. Chem.*, **42**, 1467–68 (1950).

JAMES W. PERRY AND ROBERT S. CASEY

Reprinted from:

ENCYCLOPEDIA OF CHEMICAL TECHNOLOGY

Edited by RAYMOND E. KIRK
Head, Department of Chemistry, Polytechnic Institute of Brooklyn
and DONALD F. OTHMER
Head, Department of Chemical Engineering, Polytechnic Institute of Brooklyn
Assistant Editors
JANET D. SCOTT *and* ANTHONY STANDEN

VOLUME 8
ION EXCHANGE
to
METAL PLATING

Published by
THE INTERSCIENCE ENCYCLOPEDIA, INC. • NEW YORK

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150002-0